Statistical Methods and Optimization in Data Mining

Eloísa Macedo[1], **Adelaide Freitas**[2]

[1] *University of Aveiro, Aveiro, Portugal;* macedo@ua.pt
[2] *University of Aveiro, Aveiro, Portugal;* adelaide@ua.pt

The increasing number of the sequenced genomes has created new challenges in several scientific domains, namely statistics, optimization and computer sciences. Various numerical transformations related to the sequenced genomes (e.g., frequency of each nucleotide, association between consecutive genomic symbols) have been proposed in order to take advantage of statistical methodologies available for quantitative data. It is expected that such numerical data sets contain useful information about mathematical properties of DNA sequences. An important issue associated with data sets where each individual is characterized by a high-dimensional vector of variables consists in the identification of patterns or homogeneous groups. Since high dimensionality turns the visualization and analysis of data into a complex problem, the space reduction and the features subset selection techniques are aimed to facilitate the visualization and capture the important and relevant relationships existing in data.

To detect the existence of patterns in a data matrix (n objects $\times$ p variables), it is often desirable to partition the data sets according to some similarity criteria. This task is related to the data mining technique of partitioning data sets into groups of objects with some similar properties (clusters) called clustering. There exists a variety of clustering techniques designed for several data types, applied in many areas such as pattern recognition, image segmentation and bioinformatics [1,4,5]. Clustering problems are usually formulated as mixed-integer problems, or $(0,1)$-semidefinite and semi-infinite programming problems that in turn can be reduced to nonsmooth and nonconvex nonlinear problems [2,4].

While dimensionality reduction of objects is usually achieved by clustering techniques, the dimensionality reduction of the variable space can be provided applying statistical techniques such as Principal Component Analysis (PCA), to detection of a lower number of uncorrelated variables (components) able to explain the maximum variability of the data. The reduction of objects and variables can be obtained applying the two

techniques sequentially. Recently, a new technique called Clustering and Disjoint Principal Component Analysis (CDPCA) was suggested in [3] to solve the clustering of objects and the partition of variables using PCA simultaneously. This technique permits to cluster objects along a set of centroids and to partition of variables along a reduced set of components, in order to maximize the between cluster deviance of the components in the reduced space. The model obtained is a quadratic mixed continuous and integer optimization problem. In [3], this model is solved by an alternating least-squares (ALS) algorithm that can be considered as an heuristic that divides the model solving iteratively in four steps, modifying in each step certain parameters of the data. The methods of Mixed-Integer Programming are used on the basic steps of the algorithm. In [3], the CDPCA algorithm was tested for two data sets, one with 20 objects and 6 variables and other with 103 objects and 12 variables.

The main objective of this work is to test the ability of this new technique on biological data sets to make possible visual representation of relevant characteristics for data interpretation. For this purpose, we implemented CDPCA in R language, which is an open source software widely used in statistics, with a lot of specific packages for efficient data treatment [6].

Let us introduce the notations that will be used.
— $X = [x_{ij}]$ is the data matrix with I objects and J variables (variables are supposed to be normalized);
— P, Q are the desirable numbers of clusters of objects and variables, respectively;
— E is the $I \times J$ error matrix;
— $U = [u_{ip}]$ is a $I \times P$ binary matrix and row stochastic defining a partition of objects into P clusters where $u_{ip} = 1$ if the i-th object belongs to cluster p, otherwise, $u_{ip} = 0$;
— $V = [v_{jq}]$ is a $J \times Q$ binary matrix and row stochastic defining a partition of variables into Q clusters where $v_{jq} = 1$ if the j-th variable belongs to cluster q, otherwise, $v_{jq} = 0$;
— A is the $J \times Q$ matrix of the coefficients of the linear combination, such that $rank(A) = Q$ and each row (variable) contributes to a single column (component);
— $Y = [y_{iq} = \sum_{j=1}^{J} a_{jq} x_{ij}]$ is the $I \times Q$ component score matrix where y_{jq} is the value of the i-th object for the q-th component y_q (common

information of a subset of variables);
— $\bar{X}$ is the $P \times J$ matrix of individual centroids in the space of the observed variables;
— $\bar{Y}$ is the $P \times Q$ matrix of individual centroids in the reduced space.

The model associated to CDPCA minimizes the norm of the error matrix

$$E = X - U\bar{Y}A^T$$

w.r.t. parameters representing U, $\bar{Y}$ and A subject to certain constraints. According to [3], A can be decomposed in the form $A = BV$, where B is a $J \times J$ diagonal matrix of the form

$$B = \sum_{q=1}^{Q} diag(v_q)diag(c_q),$$

where v_q is the vector corresponding to column q in matrix V and c_q is the eigenvector associated to the largest eigenvalue of the matrix

$$diag(v_q)\bar{X}^T U^T U \bar{X} diag(v_q).$$

We can formulate the problem as follows.

$$\begin{aligned}
&\max_{U,\bar{X},B,V} \|U\bar{X}BV\|^2 = \max_{v,c,\bar{x},u} \sum_{p=1}^{P}\sum_{q=1}^{Q}\left(\sum_{j=1}^{J} v_{jq}c_q\bar{x}_{pj}\right)^2 \sum_{i=1}^{I} u_{ip} \\
&\text{s. t. } \sum_{p=1}^{P} u_{ip} = 1, \quad u_{ip} \in \{0,\ 1\}, \quad i = 1, ..., I; p = 1, ..., P, \\
&\sum_{q=1}^{Q} v_{jq} = 1, \quad v_{jq} \in \{0,\ 1\}, \quad j = 1, ..., J; q = 1, ..., Q, \qquad (P) \\
&\sum_{j=1}^{J} c_{jq}^2 = 1, \quad q = 1, ..., Q, \\
&\sum_{j=1}^{J} c_{jq}c_{jr} = 0, \quad q = 1, ..., Q-1; r = q+1, ..., Q.
\end{aligned}$$

The alternating least-squares algorithm suggested in [3] alternates four basic steps: update V (allocation of variables), update B (the PCA step), update U (allocation of objects) and update $\bar{X}$ (centroid matrix), and it

is summarized in the following box. Here, the estimates of the matrices are denoted by $\hat{}$.

ALS Algorithm for CDPCA

`input:` numeric data matrix X and tolerance ε

Generate (e.g. randomly) $\hat{U}$ and $\hat{V}$, considering the constraints of problem (P);

Compute $\hat{\bar{X}} = \left(\hat{U}^T\hat{U}\right)^{-1}\hat{U}^T X$. Set k=1;

`while` $F_{k+1}(\hat{B},\hat{U},\hat{\bar{X}},\hat{V}) - F_k(\hat{B},\hat{U},\hat{\bar{X}},\hat{V}) < \varepsilon$:

Update B: Given $\hat{\bar{X}},\hat{U},\hat{V}$, calculate $\hat{B} = \sum_{q=1}^{Q} diag(v_q)diag(c_q)$.

Update V: Given $\hat{B},\hat{\bar{X}},\hat{U}$, for $j = 1,...,J$, set:

$$\hat{v}_{jq} = \begin{cases} 1, \text{ if } F(\hat{c}_q,\hat{U},\hat{\bar{X}},[v_{jq}]) = \max\limits_{r=1,...,Q}\left\{F(\hat{c}_r,\hat{U},\hat{\bar{X}},[v_{jr}=1])\right\} \\ 0, \text{ otherwise.} \end{cases}$$

where $F(\hat{c}_q,\hat{U},\hat{\bar{X}},\hat{V}) = \|\hat{U}\hat{\bar{X}}\hat{B}\hat{V}\|^2$.

Update U: Given $\hat{B},\hat{\bar{X}},\hat{V}$, for $i = 1,...I$, set:

$$\hat{u}_{ip} = \begin{cases} 1, \text{ if } \|\hat{V}^T\hat{B}x_i - \hat{V}^T\hat{B}\hat{\bar{x}}_p\|^2 = \min\limits_{s=1,...,P}\left\{\|\hat{V}^T\hat{B}x_i - \hat{V}^T\hat{B}\hat{\bar{x}}_s\|^2\right\} \\ 0, \text{ otherwise.} \end{cases}$$

Update $\bar{X}$: Given $\hat{B},\hat{U},\hat{V}$, calculate $\hat{\bar{X}} = \left(\hat{U}^T\hat{U}\right)^{-1}\hat{U}^T X$.

Compute $F_k(\hat{B},\hat{U},\hat{\bar{X}},\hat{V}) = \|\hat{U}\hat{\bar{X}}\hat{B}\hat{V}\|^2$.

`do` $k = k+1$;

The algorithm stops when the difference between consecutive computations of the values of the objective function of problem (P) is smaller than a specified threshold $\varepsilon > 0$. According to [3], since $F(B,U,\bar{X},V)$ is bounded above, the algorithm converges to a stationary point, which is a local maximum of problem (P). To guarantee that the algorithm finds the global minimum, the authors of the heuristic in [3] suggest to apply the algorithm repetitively for different initial values of matrices U and V, that are randomly chosen.

In order to test the ability of the CDPCA to reveal and visualize biologically meaningful patterns in a 2-dimensional reduced space, we have implemented the algorithm using R and carried out an experimental study involving several real data sets extracted from molecular biology domain. Besides the matrices U, V, A, the implementation of CDPCA suggested in

this work returns a pseudo-confusion matrix and draws two scatterplots where the data are displayed in the 2-dimensional reduced space, one where the objects are labelled according to the real classification and other with the classification found by CDPCA. The pseudo-confusion matrix indicates the number of objects introduced in each cluster (the real and that found by CDPCA).

On the basis of the realized numerical tests we conclude that the implementation of the CDPCA algorithm in R is efficient for the tested data sets. The main advantage of this technique is that each component is characterized by a disjoint set of variables. This offers a promising approach for the clustered visual representation of data. On the other hand, it permits to overcome the difficulties on the interpretability of the data in the reduced space. The proposed heuristic can be improved, since we can update the parameters of problem (P) simultaneously using optimization methods that efficiently use the structure and properties of this problem. This is a subject of further research.

This work was supported by *FEDER* founds through *COMPETE*–Operational Programme Factors of Competitiveness and by Portuguese founds through the *Center for Research and Development in Mathematics and Applications* (CIDMA, University of Aveiro) and the Portuguese Foundation for Science and Technology (FCT).

REFERENCES

1. *A. Freitas, V. Afreixo, M. Pinheiro, J. L. Oliveira, G. Moura, M. Santos.* "Improving the performance of the iterative signature algorithm for the identification of relevant patterns," Statistical Analysis and Data Mining, **4**, No. 1, 71–83 (2011).
2. *J. Peny, Y. Wei.* "Approximating K-means-type clustering via Semidefinite Programming," SIAM J. OPTIM., **18**, No. 1, 186–205 (2007).
3. *M. Vichi, G. Saporta.* "Clustering and Disjoint Principal Component Analysis," Computational Statistics and Data Analysis, (2008).
4. *G.-W. Weber, P. Taylan, S. Ozogur, B. Akteke-Ozturk.* "Statistical Learning and Optimization Methods in Data Mining," in Ayhan, H. O. and Batmaz, I.: Recent Advances in Statistics, Turkish Statistical Institute Press, Ankara, 2007, pp. 181–195.
5. *R. Xu, D. Wunsch.* "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, **16**, pp. 645–648 (2005).
6. *The R Project for Statistical Computing,* http://www.r-project.org/